Leveraging Open Access publishing to fight fake news

Charles Letaillieur and Sylvain Massip

The opportunity behind Open Access

In 2018, **Open Access** represented about **28%** of total published scientific articles[1]. There are many arguments in favour of Open Access to scholarly publication. In particular, it is said to improve appropriation of research by interested citizens. In this work, we study an example of a possible use of Open Access to improve citizens' information. Our main hypothesis is that **automatic information retrieval from large corpus of peer-reviewed academic research articles can help the automatic sourcing of correct answers to scientific queries from a general audience.**

Deceiving scientific claims can result in severe problems

Misleading scientific claims can be a severe problem in several domains. In particular, it can have very serious impacts in **public health** by lowering **vaccination** rates or increasing the use of **quackery medicines**.
Collectively, the published academic articles constitute the **consensus** a scientific assertion should be judged against. Hence, we try to build several **text-mining based indicators** that assess agreement between a given scientific statement and the scientific consensus.

Our objective: an integrated pipeline

The final objective of this project is to develop an integrated pipeline that would enable users to evaluate whether a scientific claim is backed by peer-reviewed literature.
We work with claims of the type:
"**Does XXX cure / cause / prevent YYY ?**"
In this exploratory study, we show in more details two indicators, built on two concrete examples.

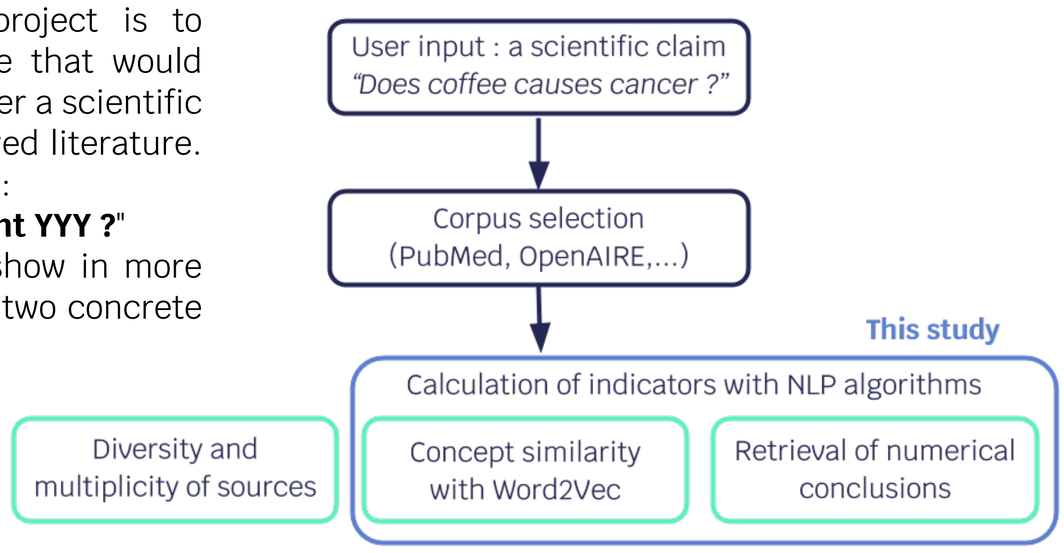

Figure 1. Pipeline architecture

Semantic similarity between concepts

DHEA-based treatments for AIDS caused serious harm in Africa[5], hence disproving it is worthwile.
A **Word2Vec**[6] model was trained on **8,233 articles** containing **"HIV"** which were retrieved from **Europe PubMed**. By measuring cosine similirity between concepts within the corpus, we show that **known treatments** (antiretroviral) are **associated with the notions of "treatment" and "therapy"** in the scientific literature whereas **DHEA** (and other hormones) are not**.**

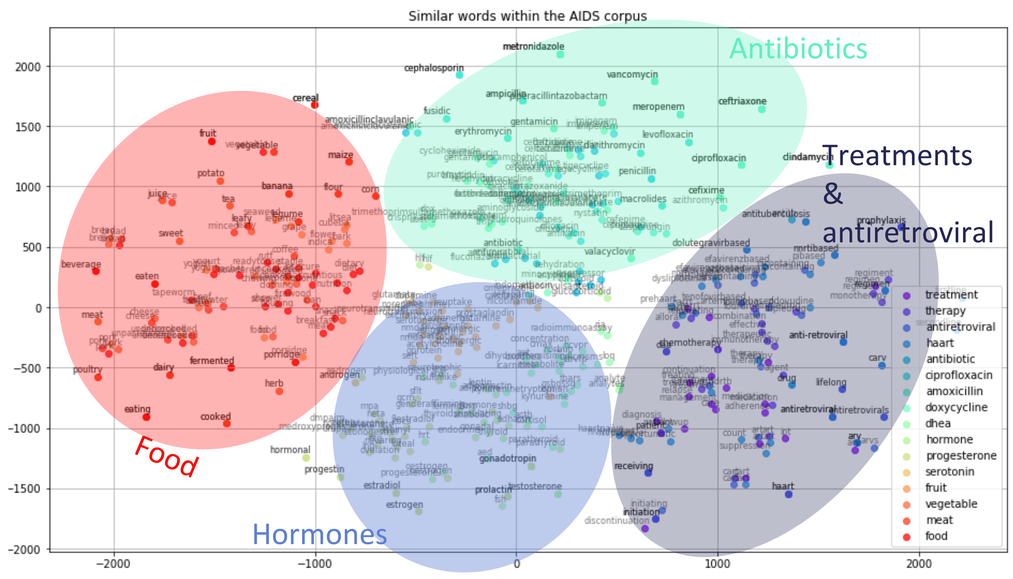

Figure 2. Similarities of concepts within an HIV related corpus

Retrieval of values

Nutrition studies often give contradictory results[2]. Hence, assessing automatically a rather consensual statement such as the link between **red meat consumption and cancer** is an interesting problem. 112 articles with "red meat" and "cancer" in their title were retrieved from Europe PubMed. **Grobid quantities**[3] was used to retrieve automatically the confidence intervals (CI) from the abstracts. A sample of 159 CI values were obtained. Their statistical distribution is shown below.

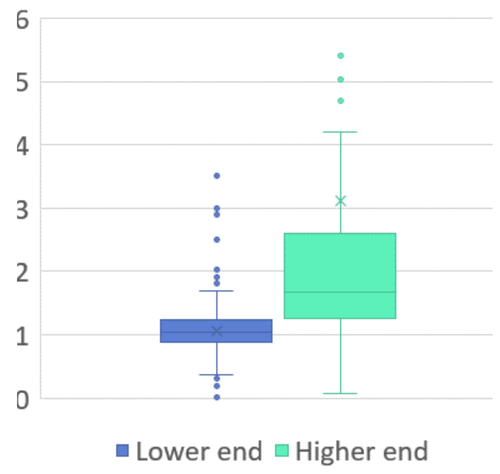

Figure 3. Automatic retrieval of confidence intervals

59 % of CI show statistical significance. Hence exhibiting a moderate consensus on the risk of red meat.
Improvements in the selection of articles, as well as in the information retrieval pipeline could allow to show a clearer consensus. The impact of potential bias, such as publication bias[4] for example, shall also be evaluated.

Conclusions

This exploratory study presents our concept of a **science fact-checker based on Open Access literature**. It further shows, on two particular examples, the use of text-mining algorithms on specialized corpus to assess whether a scientific claim is backed by the peer-reviewed literature. **Specific indicators** were built and **pipelines** to compute them automatically were developed. These results validate the feasibility of the proposed approach. The next step is to assess this methodology on a corpus of scientific claims which are **user-defined** and **expert evaluated** such as metafact.io or sciencefeedback.co. Then, an **online application** could be developed based on these principles.

Contact

www.opscidia.com
Sylvain Massip
sylvain.massip@opscidia.com
+33 6 28 30 59 20
Charles Letaillieur
charles.letaillieur@opscidia.com
+33 6 50 41 50 99

References

1. Piwowar, H., Priem, J., Larivière, V., Alperin, J. P., Matthias, L., Norlander, B., ... Haustein,S."The state of OA: a large-scale analysis of the prevalence and impact of Open Access articles." PeerJ, 6, e4375 (2018) https://doi.org/10.7717/peerj.4375
2. Jonathan D. Schoenfeld et John PA Ioannidis, "Is Everything We Eat Associated with Cancer? A Systematic Cookbook Review", The American Journal of Clinical Nutrition 97, no 1 (2013) https://doi.org/10.3945/ajcn.112.047142
3. Luca Foppiano et al., "Automatic Identification and Normalisation of Physical Measurements in Scientific Literature", in Proceedings of the ACM Symposium on Document Engineering 2019, DocEng '19 (Berlin, Germany: Association for Computing Machinery, (2019), 1–4, https://doi.org/10.1145/3342558.3345411.
4. Hannah R. Rothstein, Alexander J. Sutton, et Michael Borenstein, « Publication Bias in Meta-Analysis », in Publication Bias in Meta-Analysis (John Wiley & Sons, Ltd, 2006), 1-7, https://doi.org/10.1002/0470870168.ch1.
5. Olivier Hertel, "AIDS : Renowned Scientists Implicated in Falsified Medication Trafficking Between France and Africa", Sciences et Avenir (2016) https://www.sciencesetavenir.fr/sante/aids-renowned-scientists-implicated-in-falsified-medication-trafficking-between-france-and-africa_108869.
6. Tomas Mikolov et al., "Efficient Estimation of Word Representations in Vector Space", (2013) https://arxiv.org/abs/1301.3781